



MÉMOIRE  
AUGMENTÉE

# *L'Intelligence Artificielle au service de nouvelles opportunités pour les institutions patrimoniales*

*Futurs fantastiques – BnF – 10 décembre 2021*

ina

MÉMOIRE  
AUGMENTÉE

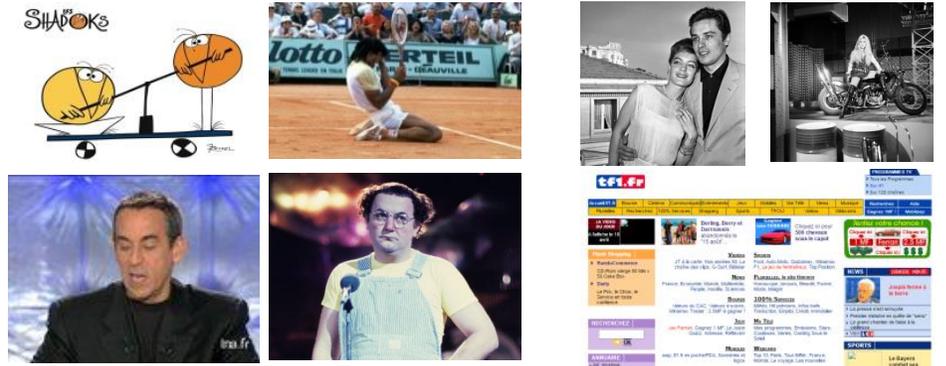
01.

→ Quelques constats

## L'Institut National de l'Audiovisuel, c'est...



100 km linéaires d'archives conservées sur supports originaux  
30 Po d'archives numérisées ou nativement numériques

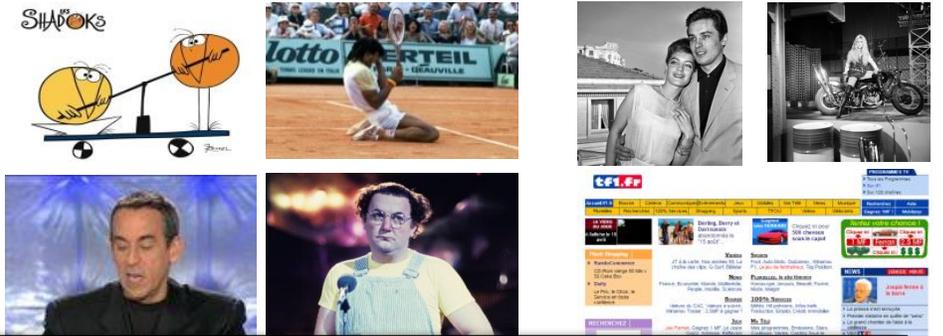


22 millions d'heures de programmes télé et radio  
2 millions de photos et 3.29 Po d'archives du Web



184 chaînes de télé/radio captées 24h/24 7j/7  
Plus de 40 millions de programmes/sujets télé/radio décrits  
pour 59 millions de diffusions référencées

## L'Institut National de l'Audiovisuel, c'est...



# Au regard des volumétries, comment exploiter au mieux la collection ?

10 30 Po d'archives numérisées ou nativement numériques

2 millions de photos et 3.29 Po d'archives du Web



184 chaînes de télé/radio captées 24h/24 7j/7  
Plus de 40 millions de programmes/sujets télé/radio décrits  
pour 59 millions de diffusions référencées

# L'INA, un média patrimonial pour...

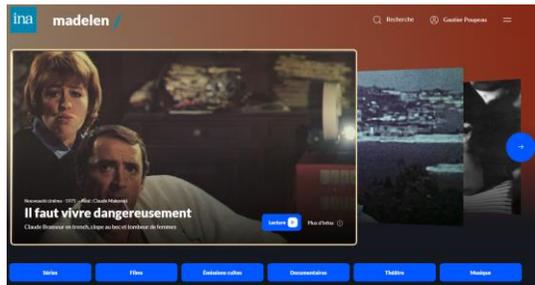
Grand public/B2C

<http://www.ina.fr>



Environ **400 000** contenus pour **80 000** heures de contenu

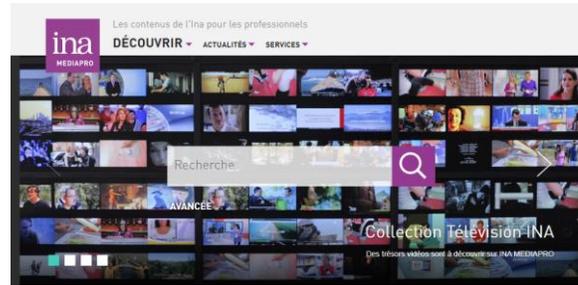
<http://madelen.ina.fr>



Catalogue SVOD d'environ **13 000** programmes

Professionnels/B2B

<https://www.inamedia.pro.com>



Près de **2 millions** d'heures disponibles

<https://mediaclip.ina.fr>



Environ **500 000** extraits prêts à l'emploi

Etudiants/chercheurs

<http://www.inatheque.fr>



**7 centres** donnent accès à tout le fonds  
**100 centres** donnent accès en autonomie à une partie du fonds

# L'INA, un média patrimonial pour...

Grand public/B2C

<http://www.ina.fr>



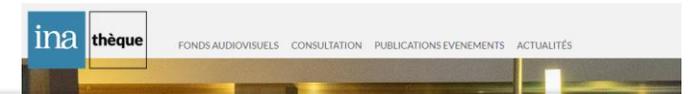
Professionnels/B2B

<https://www.inamediapro.com>



Etudiants/chercheurs

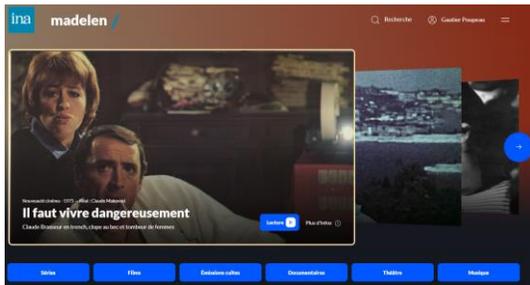
<http://www.inatheque.fr>



Mais, malgré nos efforts, ce n'est qu'une goutte d'eau dans l'océan...

80 000 heures de contenu

<http://madelen.ina.fr>



Catalogue SVOD d'environ  
13 000 programmes

Pres de 2 millions d'heures disponibles

<https://mediaclip.ina.fr>



Environ 500 000 extraits prêts à l'emploi



7 centres donnent accès à tout le fonds  
100 centres donnent accès en autonomie à une  
partie du fonds

Dans le même temps, la société interroge les médias :  
fake news, diversité, pluralisme des sujets, présence à l'écran...

**L'Union européenne dans les journaux télévisés**  
Cnews mise en demeure par le CSA pour avoir relégué LFI et le gouvernement la nuit  
PLURALISME Après cet avertissement, la chaîne promet de mieux faire à partir du 1er janvier  
20 Minutes avec AFP | Publié le 04/12/21 à 02h53 — Mis à jour le 04/12/21 à 02h54  
91 COMMENTAIRES 46 PARTAGES

**Parité hommes-femmes dans les médias : faut-il moduler les aides à la presse ?**  
Télévision. Les JT du 13h, stars du petit écran  
Les 13 h ont changé de présentateur en début d'année. Sans impact pour le moment sur ceux qui les regardent, d'abord fidèles à une chaîne et à leurs habitudes.

**Pluralisme des débats, temps de parole... Le CSA est-il loin du compte ?**  
5 minutes à lire Article réservé aux abonnés

**Zemmour sur CNews depuis un an: ce qu'il y a derrière ces audiences en flèche**  
1,6%  
Arrivé le 14 octobre 2019 dans "Face à l'Info", le chroniqueur a dopé les audiences de la chaîne. Une enquête de BFM TV. La preuve en 4 graphiques

**Femmes et médias dans le monde : les inégalités perdurent**  
ÉGALITÉ ET DROITS HUMAINS / ENDEUX DE SOCIÉTÉ /  
Valentine Ambert - Rédactrice - Youmatter  
Rédactrice pour Youmatter. Formée à Sciences Po Lyon, spécialisée sur les enjeux de développement en Afrique subsaharienne et investie dans les secteurs de la RSE, du progrès social et de la transition écologique.  
Publié le 30 novembre 2020

**LES DÉCODEURS**  
Pour comprendre En un graphique Vérification Les enquêtes des Décodeurs Datavisualisation  
Cinq idées reçues sur la chasse passée au crible : « un loisir de citadins », « moins d'accidents depuis vingt ans »  
Pays de la Zone 11 les plus vaccinés que de non vaccinés soit à l'hôpital.  
11 fois plus.  
7 arrêts les explications statistiques des pros des chiffres.  
IMPRÉCIS  
« Onze fois plus de vaccinés à l'hôpital » : comment faire mentir de vraies statistiques sur le Covid-19 ?  
Les arrêts cardiaques chez les jeunes sportifs ont-ils augmenté avec le vaccin contre le Covid-19 ?  
Faut-il vacciner les enfants de moins de 12 ans contre le Covid-19 ? Le point sur ce que l'on sait des bénéfices et des risques

**vrai ou fake**  
"Vrai ou Fake" est la plateforme de fact-checking et de debunking de l'ensemble de l'audiovisuel public. Elle rassemble des contenus produits par Arte, l'Institut national de l'audiovisuel, France 3, France 4, France 5, France 6, France 7, France 8, France 9, France 10, France 11, France 12, France 13, France 16, France 17, France 18, France 19, France 20, France 24, France 40, France 41, France 42, France 43, France 44, France 45, France 46, France 47, France 48, France 49, France 50, France 51, France 52, France 53, France 54, France 55, France 56, France 57, France 58, France 59, France 60, France 61, France 62, France 63, France 64, France 65, France 66, France 67, France 68, France 69, France 70, France 71, France 72, France 73, France 74, France 75, France 76, France 77, France 78, France 79, France 80, France 81, France 82, France 83, France 84, France 85, France 86, France 87, France 88, France 89, France 90, France 91, France 92, France 93, France 94, France 95, France 96, France 97, France 98, France 99, France 100.  
Portager Twitter Envoyer  
Les décodeurs La charte  
Les 1  
Vrai ou Fake : l'OMS "défavorable" à la troisième dose de vaccin ?

Cela doit nous amener à nous interroger....

Comment exploiter et donner du sens à cette *masse de données*,  
*au-delà* de nos missions traditionnelles de *conservation* et d'*accès* aux personnes accréditées ?

Comment *rendre tangibles* ces éléments tout en respectant nos impératifs juridiques ?

En quoi cette masse peut répondre aux *interrogations politiques et sociales* ?

Jusqu'à quel point pouvons-nous rendre cette masse *intelligible et comment* ?

Quel est *notre rôle* dans cette perspective ?

02.

→ Notre proposition

# Une rencontre entre...

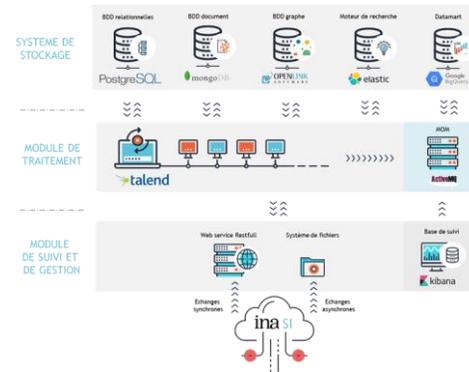
## Des corpus

issus de la captation de 184 chaînes de télé et radio pour le dépôt légal



Chaînes d'information continue

## Des technologies



Une plateforme de stockage et traitement des données en masse

## Des questions

Sur la diversité dans les médias

Comment se répartit la parole des hommes et des femmes dans les médias ?

Sur l'actualité vue par les médias

Quels sont les personnes, les lieux et les thèmes traités, cités, vus dans les médias ?

Sur le contenu des médias

Quel est l'offre de programmes dans les médias ? Comment évolue-t-elle ?



Journaux télévisés



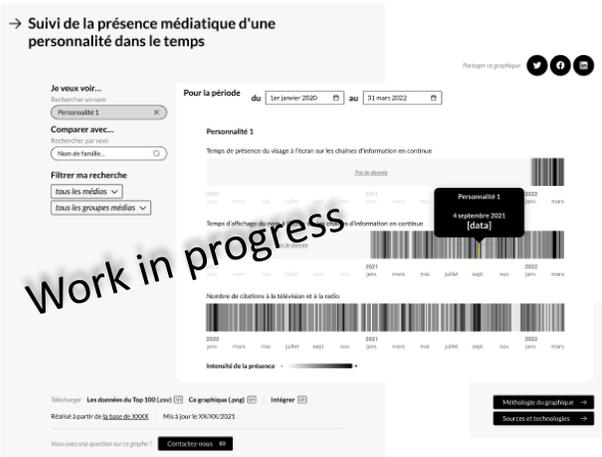
Des systèmes d'analyse automatique de l'image et du son



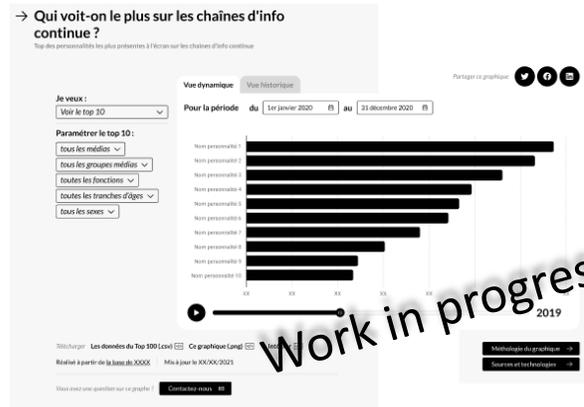
Le site sera disponible au cours du 2<sup>ème</sup> trimestre 2022 à l'adresse <https://data.ina.fr>

# ... pour proposer un site d'analyse des médias accessible à tous

Des tableaux de bord interactifs pour proposer des chiffres dans la durée sur les mêmes indicateurs



Exemple de tableau de bord sur les personnalités



Exemple de tableau de bord sur les personnalités

Les données d'indicateurs sont téléchargeables et disposent d'une licence ouverte pour en permettre la réutilisation

## Voir les données par...

Découvrez nos différentes entrées



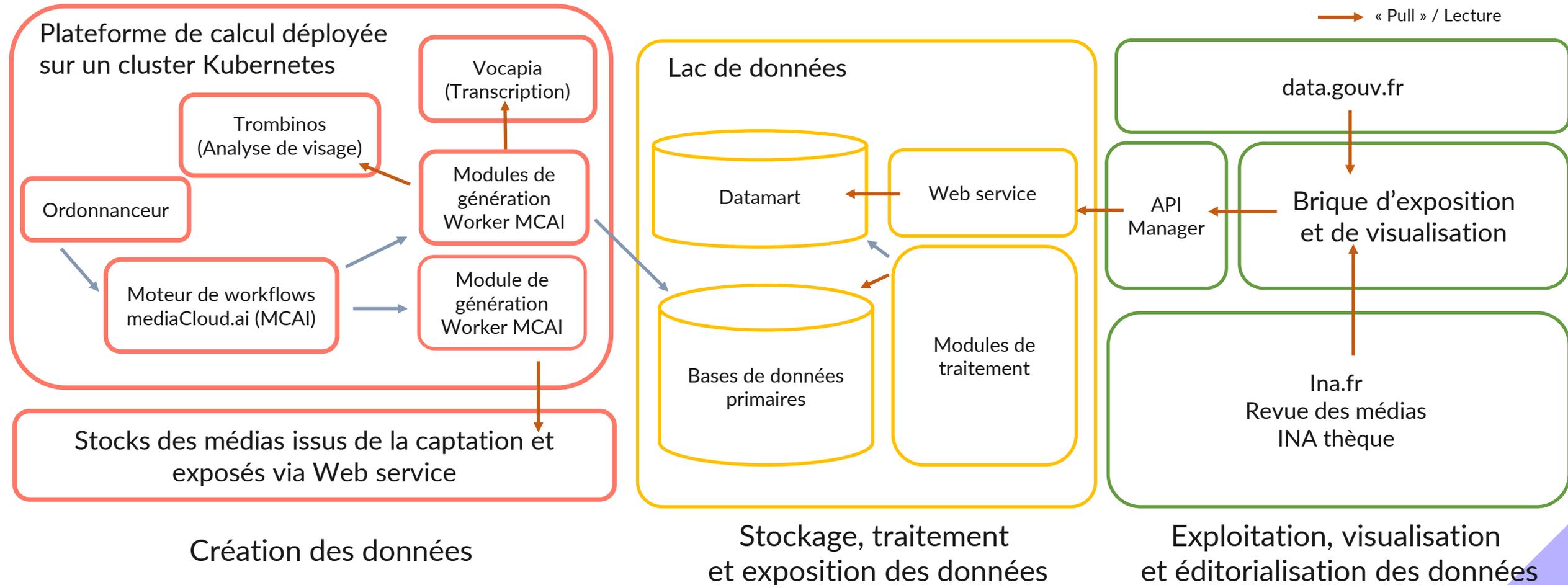
Les indicateurs sont répartis selon deux axes :

- des corpus (journaux télévisés, chaînes d'information continue et programmes pour la V1)
- des thématiques (Personnalités, diversité, lieux, mots et thèmes pour la V1)

Le travail de design, de maquettage et d'UX/UI est effectué par la société [WeDoData](https://www.wedodata.com)

# Architecture de la solution

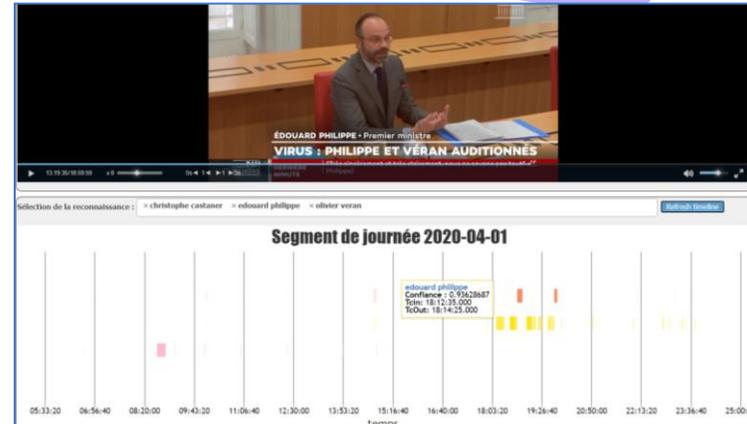
→ « Push » / Ecriture  
→ « Pull » / Lecture



# Création des données (les algos)



Transcription effectuée avec [Vocapia](#) et extraction d'entités nommées avec [TextRazor](#) et reliés à notre référentiel via un alignement avec [Wikidata](#)



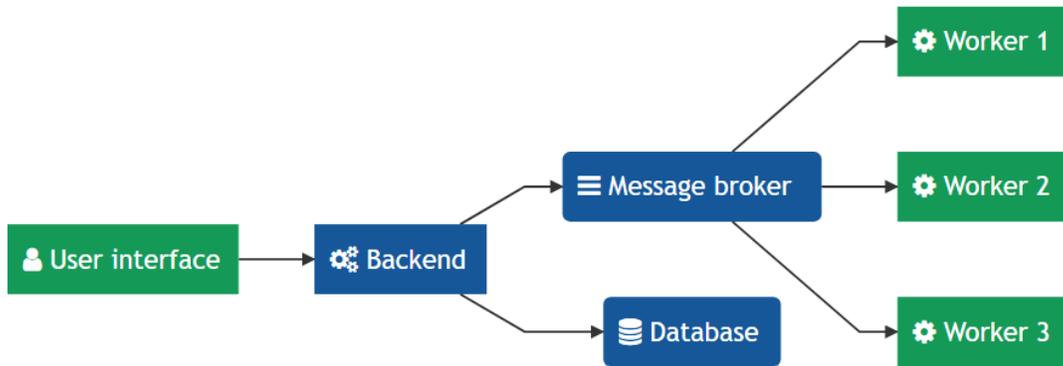
Analyse du son (homme, femme, bruit, musique) avec [InaSpeechSegmenter](#) mis au point par David Doukhan, chercheur à l'Ina



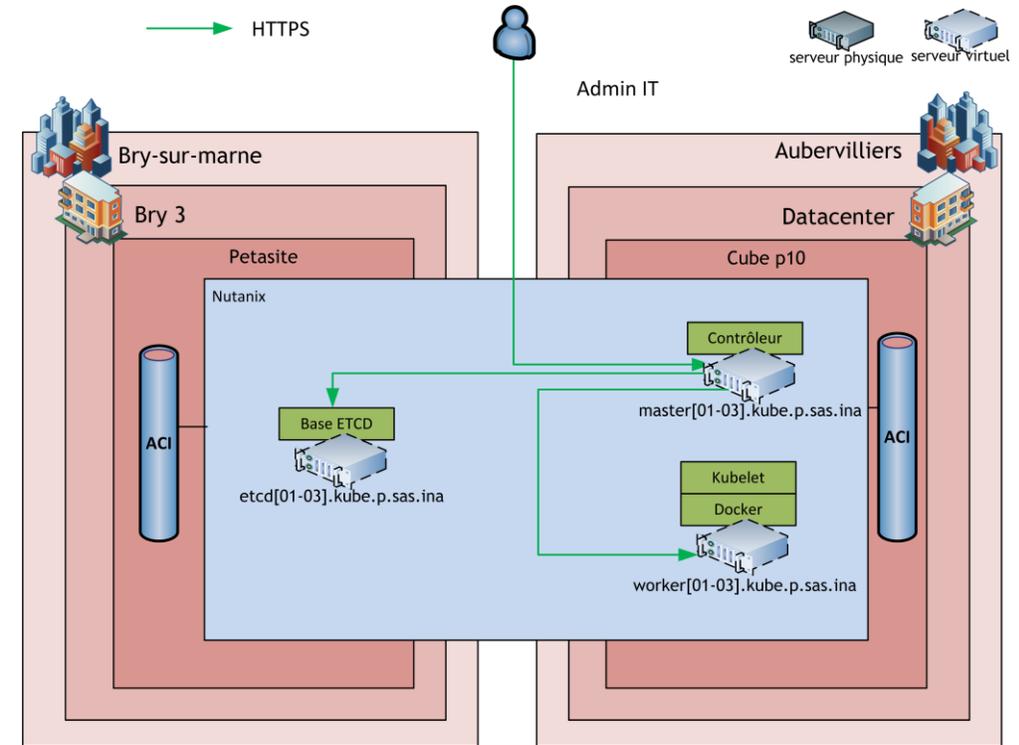
Identification automatique des visages de + 70 000 personnes avec [Trombinos](#) mis au point par Pierre Letessier, chercheur à l'INA

On peut aussi citer : l'OCR effectué avec la bibliothèque Open Source [EasyOCR](#), classification automatique des images avec la solution [Deepomatic](#), traitement automatique du langage avec les frameworks [Transformers](#) et [Tensor Flow](#), la segmentation automatique de programme, la détection de jingles, le repérage des visages et la reconnaissance d'images avec [OpenCV](#) et bien-sûr l'ensemble des métadonnées et référentiels mis au point depuis 46 ans par les professionnels de l'information de l'Ina...

# Création des données (la plateforme)

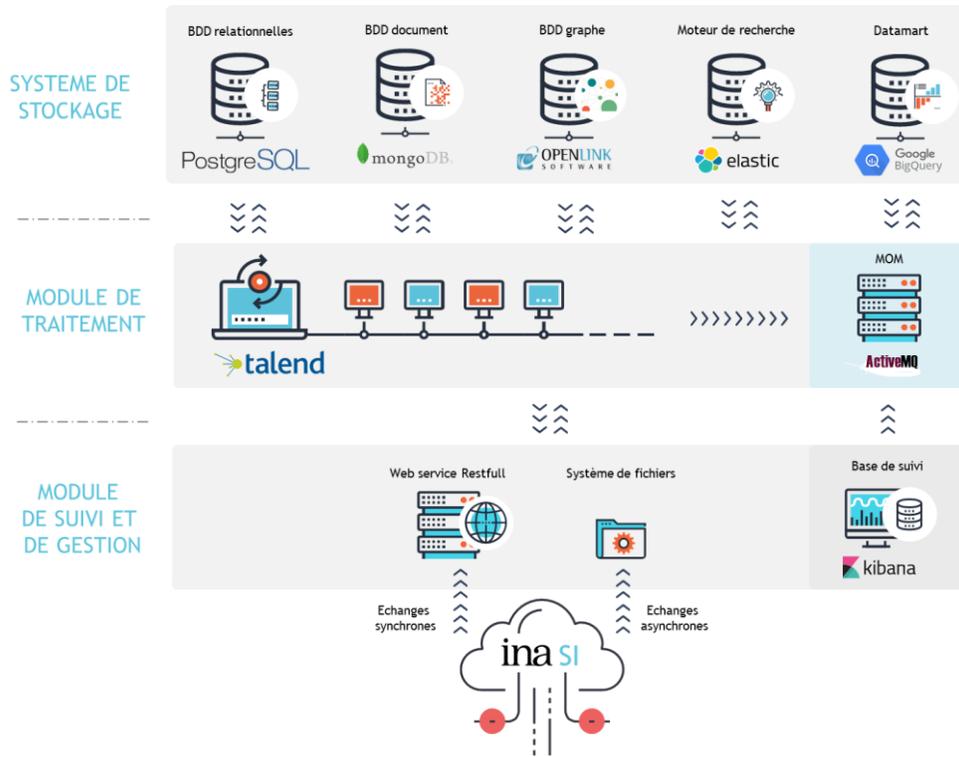


L'ensemble des traitements sont orchestrés par le moteur de workflow, [Media Cloud AI](#), plateforme Open Source de micro-services scalable et distribuée, mis au point et maintenu par France Télévision et la société Media-IO.

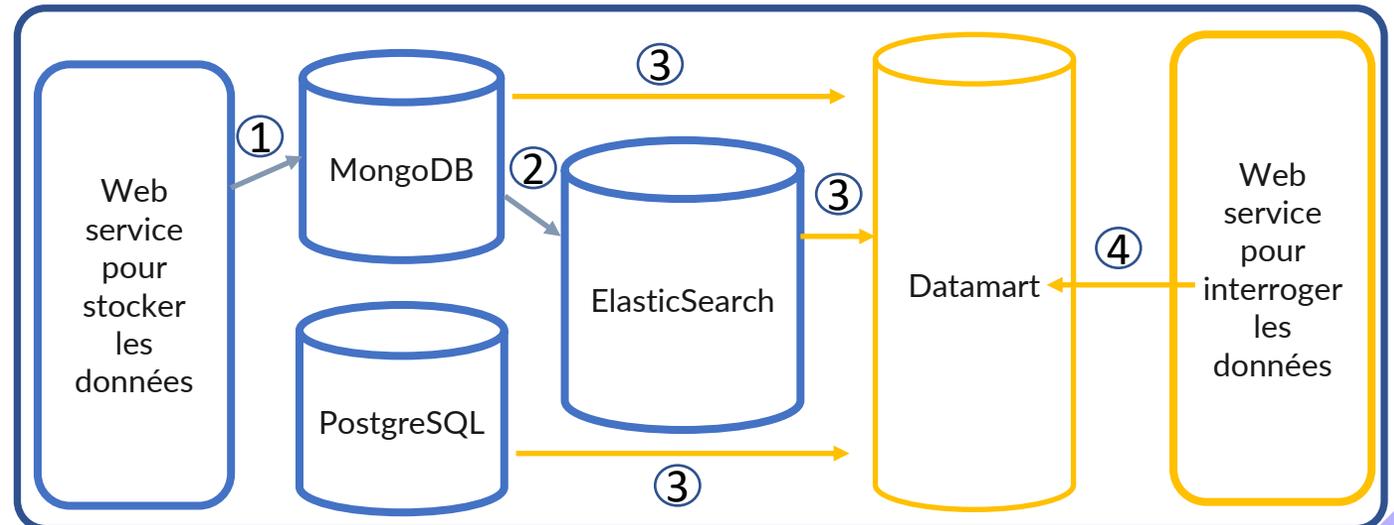


La plateforme est installée sur un cluster Kubernetes déployé sur nos propres infrastructures réparties sur deux sites physiques.

# Stockage et traitement des données



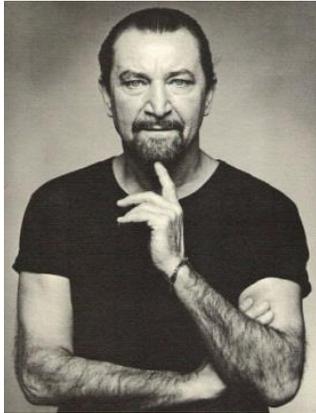
Une infrastructure pour stocker, mettre en cohérence, traiter et exposer toutes les données de l'INA



## Les différents rôles



Architecte  
de données



*Data scientist*



Ingénieur  
de la donnée



Analyste de la donnée  
Professionnels de l'information



*Data ops/Dev Ops*